DOCUMENT RESUME

ED 306 246 TM 013 018

AUTHOR Wolcott, Willa; And Others
TITLE Discrepancies in Essay Scoring.
PUB DATE Aug 88
NOTE 39p.
PUB TYPE Reports - Research/Technical (143) --
Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Students; *Essay Tests; Higher Education; *Holistic Evaluation; Interrater Reliability; *Scoring; Teacher Education Programs; *Testing Problems; Writing (Composition); *Writing Evaluation
IDENTIFIERS Comparability; *Discrepancy Analysis

ABSTRACT

The occurrence of discrepant, or non-contiguous, scores during holistic scoring of writing proficiency is usually blamed on variables of the test reader or external factors. It is hypothesized that characteristics of the essays themselves sometimes may generate discrepancies in essay scoring. Data from 152 essays of college sophomores in teacher preparation programs were used. Those essays with discrepant ratings were compared with an approximately equal number of essays with comparable scores. The same two readers read every paper to control for reader variability. Discrepantly scored essays did contain some distinguishing features that comparably scored essays did not. The use of jargon and multiple spelling errors were the few statistically significant categories contributing to discrepant scores on low-scoring papers. On high-scoring papers, discrepant scores occurred in rhetorical issues and sentence errors, English-as-a-Second-Language errors, and awkward sentences. Since the results indicate that certain characteristics of the papers themselves can contribute to discrepant scoring, it is important to train readers to be alert to potential problems. Three data tables are provided, and the scoring questionnaire is appended. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED306246

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

PATRICIA DOVELL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Discrepancies in Essay Scoring

Willa Wolcott, Patricia Dovell,
and Dianne Buhr

Office of Instructional Resources

University of FLorida
August 1988

TM 013018

Because holistic scoring has been increasingly used as a method for the large-scale assessment of students' writing proficiency, the issues of reliability and validity are critical. For this reason, any occurrence of discrepant scores--that is, non-contiguous scores--during an actual scoring session is troublesome. Not only do such discrepancies need to be refereed, thereby requiring an additional, time-consuming reading, but they also underscore potential problems with this subjective assessment approach. The reasons why such discrepancies occur are subject to debate. Most problems are attributed to the variable of the individual reader, to what Richard J. Stiggins (1982) calls "poor scoring procedures (e.g., rater leniency, halo effects, tendency toward middle rating)" (p. 110). Alternatively, discrepant scores are sometimes implicitly blamed on such scoring factors as the time of day, the order of essay distribution, or any distractions in the room--environmental conditions about which, as Sarah Freedman notes, all too little research has been done (p. 246). But another, little-studied reason may partially account for the occurrence of discrepant scores during any holistic scoring, no matter how thorough the training or how experienced the readers. The essays themselves may generate the problems; that is, the papers receiving discrepant scores may share common characteristics not typically found in those essays given either identical or contiguous scores.

Literature Review

To be certain, many researchers have already attempted to identify those factors that can influence the response of raters. For example, in their careful study of the criteria used by readers holistically scoring timed essays, Hunter Breland and Robert Jones (1984) found that organization,

development, and quality of thought influenced readers the most, with length being the most significant textual feature. Similarly, the importance of length in writing judgments, as well as correct spelling, was noted by Murray Stewart and Cary Grobe when they studied the impact of syntactic maturity and mechanics on teachers' holistic evaluations of expository essays composed by 5th, 8th and 11th grade students; these factors also appeared significant in Grobe's subsequent adaptation of the earlier study. When Ellen Nold and Sarah Freedman (1977) explored which variables affected the response of readers to 88 essays written by Stanford freshmen, they, too, found length to be correlated with quality. A study by Bennett Rafoth and Donald Rubin (1984) in which the content of an expository essay was manipulated found that mechanics, especially spelling, interfered with scorers' independent evaluations of content in that essay, even when some graders were specifically advised to ignore mechanics; however, the scoring was not a controlled holistic reading.

Still other researchers have attempted to determine the effect that external or surface features can have on writing judgments. For example, in Clinton Chase's (1968) study of variables affecting test scores, Chase found that papers with good handwriting tended to receive better scores than did papers with poor writing. Comparable findings occurred in Jon Marshall's study (1972) in which he noted that grades given to handwritten versions of papers containing up to six errors were lower than were grades given to typed versions of the same papers. The negative influence of handwriting is further implied in the studies of Diederich, French, and Carlton (1961) and of Nold and Freedman (1977) in which papers were typed so that handwriting would <u>not</u> be an issue. In William McColly's study (1970), both handwriting and external

appearance of essays were found to be such strong contributing factors that McColly recommended that "examination essays [be] typed or put into some other standard printed format" (p. 154).

A recent effort to determine what causes "variant judgments" in the holistic scoring of essays was undertaken by Loren Barritt, Patricia Stock, and Francelia Clark (1986) in which faculty gathered twice a month to read together a placement essay, to make an individual judgment about its quality, and to note the reasons for this rating.

After two years of inquiry, the researchers found that the comments fell into three general categories--those that dealt with the written product, those concerned with the author behind it, and those that hypothesized about the author as a prospective student. Barritt, Stock and Clark offered the following thoughtful explanation about this discrepancy.

> ...in responding to essays that resembled those of the majority of the eighteen-year-old college freshmen we had come to know, we reader/evaluators seemed to focus on the text itself. But even in these cases we believe the apparently exclusive focus on the written text to be a misleading indication of the basis of readers' judgments and evaluations: in conventional essays, a reader/evaluator implicitly judges that the 'student' is the one expected: a 'typical' eighteen-year-old college freshman. When an essay matches our expectations, the student as author goes otherwise unnoticed. In essays . . . when the expected student is not found, the reader begins her active characterization of the author so that she can work together with the student writer to construct a coherent text." (321-322)

For Barritt, Stock, and Clark, then, it is the active participation of the readers that leads to the occasional--and justifiable--discrepancy in ratings. Arguing that "...there is no good reason why two different readings might not both be accurate ones" (p. 324), they note that those essays which produce discrepant ratings may, in fact, be the most instructive. Thus, Barritt, Stock, and Clark suggest that evaluators focus not so much on

obtaining better agreement, but rather on trying to understand the reasons behind the inconsistencies. Their view seems echoed by Davida Charney, who believes that the unnatural reading speed of holistic scoring, as well as its pressure on readers to conform, may deny scorers the chance to respond thoughtfully to writing (p. 74).

Hypotheses and Rationale

Precisely in order to understand why scorers respond the way they do in some contexts, this study replicated an earlier one to determine whether certain features might distinguish papers receiving discrepant scores from those that received comparable scores.

A preliminary study completed last year suggested that certain characteristics within the essays do, in fact, appear to distinguish the papers receiving identical or contiguous ratings from those receiving discrepant ratings. Several distinguishing characteristics became evident when the papers being compared were subdivided into groups of high and low papers. (On a summed scale of 2 to 8, high papers had a total score of 5-8, and low papers had a total score of 2-4.) However, because the sample for that study came from only one large-scale assessment, a replication study using a sample from a different assessment situation was desirable before any generalizations could be drawn.

In accordance with the preliminary study, the same null hypothesis was used:

> Essays which receive non-contiguous or discrepant scores in an holistic evaluation do not contain any distinguishing features from those essays which receive comparable scores.

In addition, the following two ancillary hypotheses were formulated:

> Essays which receive low discrepant scores in an holistic scoring do not contain different characteristics from those essays which receive comparable low scores.

Essays which receive high discrepant scores in an holistic scoring do not contain different characteristics from those essays which receive comparable high scores.

Data Source

Data from a large-scale writing assessment were studied for the replication. The sample of essays was selected from an examination which was administered in the summer of 1987 to candidates applying for teacher certification in the state of Florida. Each examinee had 45 minutes to write an essay on one of two topics. The scoring criteria were provided to the examinees. For the preliminary study the examinees were college students at the end of their sophomore year; in that exam each student was given 50 minutes to write an essay on one of two topics comprising a different format from that of the teacher examination. Although examinees in both studies were writing essays for state-mandated competency purposes, the population of each assessment situation was thus different in terms of the number of college years completed and, to a lesser extent, in terms of the age of the examinees.

The demands of the writing tasks differed for the two assessment populations as well. The college sophomores in the preliminary study were given two non-personal, kernel topics from which to choose; both topics required students to generate a thesis, explain a position, and use their general knowledge to develop the essay. A typical topic might be "Important skills for the beginning college student." The teacher candidates, on the other hand, were given topics that contained a brief context. In addition, the teacher topics often asked for personal experiences as a means of support. For example, a typical topic might read, "The telephone has almost become a necessity in our society. What effects has the telephone had on your life?"

Thus, the teacher topics frequently called for narrative, as well as expository, responses.

In both assessments an holistic scoring approach was used. The readers were assembled on-site in one location. Readers were trained before and during the readings, and each essay was read twice. For both assessments, readers used model papers called rangefinders as their point of reference in assigning scores; in the assessment of college sophomores used for the preliminary study, operational definitions describing each score level were provided as well. For both assessments, scores of one (low) through four (high) were assigned. Each essay was read twice. The reported score is the sum of the two ratings (2-8). However, discrepant ratings (i.e. 1-3, 2-4, and 1-4) were refereed, and one of the two scores was replaced. For both assessments, the minimum passing score was 4. Also total scores of 3 were refereed. This policy was established to correct the situation in which an essay was "failed" by one rater (score of 1) and "passed" by another (score of 2).

Design and Procedures

Two groups of essays were selected for comparison. One group represented those that had received discrepant ratings (i.e., 1-3, 2-4, and 1-4). The second group represented those that had received comparable ratings (i.e., identical or contiguous scores). This study used the total population of essays with discrepant ratings, since there were relatively few papers in this category. An approximately equal number of essays with comparable ratings were selected by use of a random numbers table. High and low papers were included in each group. A high paper was defined as one having received

an actual score of 5, 6, 7, or 8 in June 1987, whereas a low paper was defined as one having received an actual score of 2 or 4.

One hundred sixty four total essays were selected. Twelve were used in training and 152 were rated according to the questionnaire. The following chart displays the sample size:

	Low	High	
Essays with Discrepant Ratings	20	52	72
Essays with Comparable Ratings	19	61	80
	39	113	

The sample size was selected to provide adequate statistical power for the chi-square analysis without placing too great a burden upon the readers.

Several steps were taken to control for reader reliability. First, the papers used in this study were scored at one site. This one-site scoring process ensured that all readers received the same training. Secondly, the two scorers who were chosen for this study to rate the papers according to the questionnaire have extensive experience in writing assessment and teaching composition. The papers and questionnaires were distributed to the raters in six weekly sessions. Each session required that training papers be rated according to the questionnaire. The purpose of these sessions was to be sure the raters stayed "on target" when rating the papers. The readers rated each set of papers during the week at their convenience, and returned them at the next training session. The readers delivered their final set of papers and comments to the researchers at a seventh meeting.

To control for reader variability, the same two readers read every paper. The readers completed the questionnaire (attached as Appendix A) for each paper without knowing either the group membership of the paper or the ratings

previously assigned. The questionnaire had been pilot-tested by the chief reader of a large writing assessment program and an associate chief reader. The questionnaire contained 27 statements--either positive or negative--to which the scorers responded by checking the appropriate category from a set marked "strongly disagree" (1), "disagree" (2), "agree" (3), or "strongly agree" (4) or alternately, from a set marked "never" (1), "seldom" (2), "occasionally" (3), or "frequently" (4). In addition the scorers could note examples of a specific problem.

The statements fell into four main categories. The first category (six items) addressed such diverse issues as the potential for controversy in the individual treatment of a topic or the relationship between content and form in a given paper. Two examples are cited below:

a) The paper reflects a comparable degree of control over rhetorical elements and grammar or usage.

b) The paper contains some special rhetorical devices (e.g., questions). If you agree, specify the device.

The second category, containing 12 items, concerned issues of rhetoric, diction, and style as noted below:

c) The paper contains a stated or clearly implied thesis.

d) The paper is focused and coherent.

e) The paper contains misinformation.

The third category, with seven items, dealt with issues of grammar, usage, or mechanics. Examples are cited below:

f) Sentence errors (e.g., fragments, run-ons, faulty parallelism, or dangling modifiers) occur in the paper. If any, specify type(s) of error.

g) ESL errors with articles, prepositions, or idioms occur in this paper. If any, specify type(s) of error.

The fourth category contained two items concerned with surface features;

h) The paper is unusually messy, or the handwriting is hard to read.

i) The length of the paper seems adequate.

The questionnaire culminated with an item that asked the scorers to check the one category best describing the paper; it contained such descriptors as "marginal," "routine and mechanical," or "smooth and solid."

Reader reliability was measured by computing the percentage of agreement between the two readers. The overall reliability was .83. The reliability by category of items on the questionnaire was as follows:

Category I (Q 1- 6) = .76

Category II (Q 7-18) = .81

Category III (Q 19-25) = .77

Category IV (Q 26-27) = .88

Validity was measured by computing the percentage of agreement between the readers and the standard. The standard was set by two chief readers of the assessment program who rated 16 papers in this study on the questionnaire. Four were used in training and twelve were used to compute the percentage of agreement. The overall validity was .75. The validity by category of items on the questionnaire was as follows:

Category I (Q 1- 6) = .61

Category II (Q 7-18) = .77

Category III (Q 19-25) = .79

Category IV (Q 26-27) = .88

The validity measure for Category I is low, and Category I also had the lowest reliability measure of the four categories. The potential reason for these findings is that Category I is the most subjective. This category contained questions open to different interpretations (such as creativity, controversial

treatment of topic, and troublesome features). Despite the training, those subjective terms gave more latitude to the scorers and hence, lowered the measures of validity and reliability.

Results

Contingency tables were prepared to describe the data collected from the questionnaires. Two sets of comparisons were made: Low discrepant ratings were compared with low comparable ratings, and high discrepant ratings were compared with high comparable ratings (Table 1). In order to report the data more succinctly the categories "agree" and "strongly agree" were collapsed in the category "agree" and "disagree and "strongly disagree" were merged as "disagree." Also the categories "never" and "seldom" were grouped together, as were "occasionally" and "frequently."

Secondly, a chi-square analysis was conducted to determine if the differences between the groups were statistically significant at the .05 level for each test. The chi-square test was selected because it is an appropriate multivariate analysis for categorical variables. The alpha level of .05 was chosen as it is a typical significance level used in educational research. Each question was analyzed with three comparisons (the overall group of discrepant with comparable ratings, and the low and high sub-groups). The statistically significant results are reported in tables 2 and 3. These tables show the effects that were significant at an alpha level of .05 or less. The chart below summarizes the data reported in table 3 and also compares the significant differences between low and high discrepant/comparable papers.

Comparison of Significant Differences Between Low and High Discrepant/Comparable Papers

		Low Group	High Group
Category I			
None			
Category II			
Q 8.	Adequately Developed	*	*
Q 9.	Organized Paper		*
Q 10.	Focused and Coherent Paper		**
Q 15.	Precise Diction		**
Q 17.	Inflated Diction/Jargon	*	
Category III			
Q 19.	Awkward Sentence Constructions		*
Q 21.	Sentence Errors		**
Q 23.	ESL Errors		*
Q 24.	Spelling Errors	*	
Category IV			
None			

* $P \leq .05$
** $P \leq .01$

As explained by Everitt (1979), the chi-square test is a test of independence. If the rater's agreement or disagreement with a statement is independent of the group membership of the essays (discrepant or comparable ratings), the proportion of essays with discrepant ratings that elicited an "agree" response would be expected to equal the proportion of those with comparable ratings that elicited the same response. If these proportions differ, then agreement with a particular statement tends to be associated more with one of the groups than with the other. Although the differences could be attributed to chance, what the researcher seeks to determine is whether the difference between proportions is too large to be attributed to that factor.

One word of caution needs to be introduced regarding the results of the chi-square analysis. The observations include two readers rating each essay on 27 separate questions. This procedure could reduce the independence of the observations. However, reader reliability research strongly supports the need for multiple ratings per essay, so the researchers believed that this design was a necessary condition of the study.

Discussion

Certain patterns appeared in the group of papers as a whole, although more information is revealed when the essays are broken into subsets of high and low scores. In the overall set of papers, the absence of certain positive features, such as coherence and focus or precise diction, contributed to discrepant scores. In contrast, the presence of certain negative features in the entire set, such as frequent spelling errors, punctuation errors, sentence errors, or inflated language, also distinguished the discrepant papers from the comparable papers at a statistically significant level. As will be seen in the following discussion of subsets, these patterns were manifested in either the high or low subset of papers; that is, the only feature which contributed in both subsets to discrepant scores--adequate development--did not appear in the overall group. Presumably, this feature was working in opposing directions in the subsets: The presence of solid development in low papers would have been surprising, whereas its absence in high papers, where it is typically expected, would have been troublesome. Precisely because the results for the overall group can mask what actually has occurred, it is essential to examine the patterns within the subsets of essays.

For example, three features appeared to distinguish at a statistically significant level the low set of discrepantly scored papers from the low set of comparably scored papers: (1) adequate development (question 8), (2) jargon or inflated diction (question 17), and spelling errors (question 24).

Almost twice as many (60%) of the ratings assigned to the discrepant set as to the comparable set (32%) indicated that the essays were adequately developed. Because most low papers are not well developed, this finding may help to account for the discrepant scores. Similar findings occurred in the first study in which more of the low discrepant papers were perceived as well developed than were the low comparable papers.

In the current study, jargon or inflated language appeared to be problematic in 12% of the low discrepant papers, whereas it did not occur as an issue at all in the high papers. Significantly, it was not problematic for any of the students in the first study either. However, because the examinees in the current study were teachers, the inappropriate use of educational terminology undoubtedly contributed to the discrepant scores.

Spelling errors were rated as occasional or frequent in nearly all (90%) of the discrepant set, in contrast to 71% of the ratings given the comparable set. In the earlier study spelling was not a distinguishing feature of the discrepant low papers; its current significance may possibly be attributed to the importance scorers attach to the examinees' ability to spell correctly in their classrooms.

With the exception of these three features, no notable characteristics appeared to distinguish the discrepant low papers from the comparable low papers. In contrast to the low set, several more features distinguished at a statistically significant level the high set of discrepantly scored papers

from the high set of comparably scored papers: (1) adequate development (Question 8). (2) adequate organization (Question 9), (3) focused and coherent papers (Question 10), (4) precise diction (Question 15), (5) awkward sentence constructions (Question 19), (6) sentence errors (Question 21), (7) ESL errors (Question 23).

One-third of the discrepantly scored papers were rated as not being adequately developed, focused, and coherent. Furthermore, almost 20% received negative ratings for organization and for precise diction. As these features are associated with good writing, their absence in upper-half papers might have contributed to the discrepant scores.

Additional problems possibly contributing to discrepant scores include awkward sentence constructions, sentence errors, or ESL errors. Over one-third of the discrepantly scored papers (38%) contained occasional or frequent use of awkward sentence constructions, and one-third (34%) contained occasional or frequent sentence errors. In addition, 10% of the papers in this set contained occasional or frequent ESL errors. High papers do not typically contain many serious sentence errors or awkward constructions; thus, their frequent occurrence in these papers might explain some of the difficulty encountered by scorers, especially if the same papers demonstrated positive characteristics in other areas. In the earlier study, the presence of awkward constructions and usage errors also contributed to the discrepant papers; in addition, in the preceding study, misinformation marred the high papers.

Summary and Discussion of Results

The original null hypothesis, along with the ancillary hypotheses, must be rejected, and it can be concluded that discrepantly scored essays do contain some distinguishing features that comparably scored essays do not

have. With the exception of adequate development, the distinguishing features which characterize low papers differ from those which characterize the high papers.

For the low papers, the only other statistically significant differences occurred in the prevalence both of jargon and of multiple spelling problems in the discrepant papers. That so few other categories contributed to discrepant scores on the low papers is interesting in view of the large number of differences which appeared in the low papers of the previous study. In the first study, for example, such categories as depth of thought, adequate conclusion, or special rhetorical devices distinguished low discrepant papers from comparable papers; in addition, problems with punctuation/capitalization and messiness/illegibility appeared as distinct problems in the low discrepant papers of the first study.

One explanation for the different results between the first study and the current work is the smaller sample size of the low papers in the present study: Whereas the early study contained a total of 148 low ratings or 74 papers altogether, the current study contained only half as many (78 ratings on 39 papers). The difference in testing populations may have also been a factor, as the teacher certification candidates were typically older than the examinees in the first study, and many had undoubtedly been teaching in other states. Additionally, unlike the other examinees, the teacher candidates were given topics that were structured and that enabled them to draw upon personal experiences if they wished; these factors may possibly have contributed to a better overall performance than was reflected in the papers used for the earlier study.

In the high papers of the current study, discrepant scores occurred in two major categories. Category II, which dealt with rhetorical issues, was especially troublesome, as the high comparable papers differed from the discrepant essays in terms of development, organization, focus, and diction. Similarly, the discrepant and comparable papers in the high set differed significantly in the Category III features of sentence errors, ESL errors, and awkward sentences. Obviously, problems with any of these areas can contribute to scorers' difficulties in assigning upper-half scores to papers.

As was found in the previous study, then, certain characteristics within the essays appear to distinguish either the high comparable papers from the high discrepant papers or the low comparable papers from the low discrepant papers. What is important to note, however, is that the nature of these characteristics seems closely allied to the particular test and may be a function of the topic, the testing population, or possibly the different criteria for the holistic scores assigned.

Certainly, factors other than the characteristics of the essays themselves may contribute to discrepant scores. For example, some researchers such as Barritt, Stock, and Clark (1986), have found that readers can, in some instances, be responding to the writer behind the work rather than to the written text. In a similar vein, Sullivan (1986) suggests that a reader may be actively constructing a writer from the amount of information contained in a paper and the degree of assumed familiarity on the writer's part that such information reflects. The viewpoints of Sullivan, Barritt, Stock, and Clark are underscored by reader response criticism, which has shown that readers bring to a text individual perspectives that can affect the interpretations of the text.

Notwithstanding the possibilities of other explanations, this study suggests that the written texts--the essays themselves--may generate discrepant scores as the presence or absence of particular features contributes to scorers' difficulty in assessing writing quality. Because the results of this study were comparable to those in the previous study, it is important to take certain steps in the training sessions of a scoring to minimize the likelihood of discrepant scores developing. Readers can continue to be alerted to potential problems with illegibility or with length; during training sessions, they can practice with sample essays that illustrate discrepancies between form and content. Similarly, one or two essays with ESL errors can be included so that scorers can consider how best to evaluate them.

Discrepancies will never entirely disappear from a scoring, but ongoing measures to minimize the differences are essential. To strive for a consensus in holistic scoring is neither to disparage the individuality of the scorers' responses nor to minimize the complexity that writing assessment of this kind entails. Rather, such efforts to obtain agreement among readers ensure that a given paper will be evaluated, insofar as humanly possible, against the same standards as any other paper, no matter who the scorer should be. Thus, the efforts to reduce discrepant scoring lessen the possibility that readers may be over-rewarding or over-penalizing certain features instead of evaluating the essay in its entirety. Precisely because reading and writing are, according to Edward White, "intensely individual activities involving the full self" (p. 32), so, too, should writing evaluation reflect a concern with the whole.

BIBLIOGRAPHY

Barritt, Loren, Stock, Patricia L., & Clark, Francelia (1986). Researching Practice: Evaluating Assessment Essays. College Composition and Communication, 37, 315-327.

Breland, Hunter M. & Jones, Robert J. (1984). Perceptions of writing skills. Written Communication, I, 101-119.

Charney, Davida (1984). The validity of using holistic scoring to evaluate writing: A critical overview. Research in the Teaching of English, 18, 65-81.

Chase, Clinton (1968). The impact of some obvious variables on essay test scores. Journal of Educational Measurement, 5, 35-38.

Cooper, Charles R., Ed. (1985). Researching response to literature and the teaching of literature: Points of departure. Norwood, NJ: Ablex Publishing Company

Diederich, B., French, S. W., & Carlton, S. T. (1961). Factors in judgements of writing ability: Research Bulletin RB 61-15, Educational Testing Service. ERIC ED002172.

Everitt, B. S. (1979). The analysis of contingency tables. London: Chapman and Hall Ltd.

Freedman, Sarah (1979). How characteristics of student essays influence teachers' evaluations. Journal of Educational Psychology, 71, 328-338.

Freedman, Sarah (1981). Influences on evaluators of expository essays: Beyond the text. Research in the Teaching of English, 15, 245-255.

Grobe, Cary (1981). Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. Research in the Teaching of English, 15, 75-85.

Marshall, Jon C. (1972). Writing neatness, composition errors, and essay grades re-examined. The Journal of Educational Research, 65, 213-215.

McColly, William (1970). What does educational research say about the judging of writing ability? The Journal of Educational Research, 64, 148-155.

Nold, Ellen W. & Freedman, Sarah W. (1977). An analysis of readers' responses to essays. Research in the Teaching of English, 13, 205.

Rafoth, Bennett and Rubin, Donald (1984). The impact of content and mechanics on judgments of writing quality. Written Communication, I, 446-457.

Steward, Murray & Grobe, Cary (1979). Syntactic maturity, mechanics of writing, and teachers' quality ratings. Research in the Teaching of English, 13, 207-215.

Stiggins, Richard J. (1982). A comparison of direct and indirect writing assessment methods. Research in the Teaching of English, 11, 164-174.

Sullivan, Francis (1986). Placing texts, placing writers: Sources of readers' judgments in university placement testing. Research/Technical Report 143. ERIC ED 285 177.

White, Edward M. (1986). Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance. San Francisco: Jossey-Bass.

APPENDIX A

Questionnaire

Analysis of Essays for The Study of Split Scores

Directions: Please assign a holistic score to each paper, using a scale of 1 to 4 with 4 the highest. Then check the category that best reflects your response to each item listed below.

Paper Number _____ Rater Code _____ Holistic Score _____

I

	Strongly Disagree 1	Disagree 2	Agree 3	Strongly Agree 4
1) The paper reflects a comparable degree of control over rhetorical elements and grammar or usage.				
2) A discrepancy exists in the paper between content and grammar. If you agree, specify the problem. _____				
3) One feature appears especially troublesome in this paper. If you agree, specify the feature. _____				
4) The paper reflects a certain degree of creativity.				
5) The paper contains some special rhetorical devices (e.g., questions). If you agree, specify the device. _____				
6) The treatment of the topic may be potentially controversial.				

II

7) The paper contains a stated or clearly implied thesis.				
8) The paper is adequately developed with specifics.				
9) The paper is organized.				
10) The paper is focused and coherent.				
11) The paper reflects some depth of thought.				

	Strongly Disagree 1	Disagree 2	Agree 3	Strongly Agree 4
12) The paper contains misinformation.				
The paper has an adequate				
13) introduction.				
14) conclusion.				
The diction is generally				
15) precise				
16) mature.				
17) inflated/jargon				
18) The sentence style is varied and mature.				

III

	Never 1	Seldom 2	Occa-sionally 3	Fre-quently 4
19) Awkward sentence constructions occur in the paper.				
20) Usage errors (e.g., pronoun reference or subject-verb agreement) occur in the paper. If any, specify type(s) of error. _____				
21) Sentence errors (e.g., fragments, run-ons, faulty parallelism, or dangling modifiers) occur in the paper. If any, specify type(s) of error. _____				
22) Dialect errors with third person verbs, noun plurals, or past participles occur in the paper. If any, specify type(s) of error. _____				
23) ESL errors with articles, prepositions, or idioms occur frequently in this paper. If any, specify type(s) of error.				

	Strongly Disagree 1	Disagree 2	Agree 3	Strongly Agree 4
24) Spelling errors occur in this paper.				

If any, check the predominant error:

_____ a) key word(s) misspelled.
_____ b) simple word(s) misspelled.
_____ c) troublesome word(s) misspelled.

25) Punctuation or capitalization errors occur in this paper.				

If any, check predominant type of error:

_____ a) Commas
_____ b) Semi-colons/colons
_____ c) Quotes
_____ d) End punctuation
_____ e) Apostrophes
_____ f) Capitalization

IV

	Strongly Disagree 1	Disagree 2	Agree 3	Strongly Agree 4
26) The paper is unusually messy, or the handwriting is hard to read.				
27) The length of the paper seems adequate.				

28) Check the category that best describes the paper.

_____ a) extremely poor
_____ b) marginal
_____ c) adequate
_____ d) stiff and overly formal
_____ e) routine and mechanical
_____ f) competent but no real flair
_____ g) smooth and solid
_____ h) outstanding

APPENDIX B

Tables

Table 1

Data Comparing Essays With Discrepant and Comparable Scores

	LOW GROUP		HIGH GROUP	
Q1. Control	**Discrepant**	**Comparable**	**Discrepant**	**Comparable**
Agree	98%	95%	99%	96%
Disagree	2%	5%	1%	4%
	100%	100%	100%	100%
	(40)	(38)	(104)	(122)
Q2. Discrepancy Between Content and Grammar	**Discrepant**	**Comparable**	**Discrepant**	**Comparable**
Agree	0%	3%	2%	5%
Disagree	100%	97%	98%	95%
	100%	100%	100%	100%
	(40)	(38)	(104)	(122)
Q3. Troublesome Feature	**Discrepant**	**Comparable**	**Discrepant**	**Comparable**
Agree	38%	39%	29%	32%
Disagree	62%	61%	71%	68%
	100%	100%	100%	100%
	(40)	(38)	(104)	(122)
Q4. Creativity	**Discrepant**	**Comparable**	**Discrepant**	**Comparable**
Agree	48%	42%	68%	70%
Disagree	52%	58%	32%	30%
	100%	100%	100%	100%
	(40)	(38)	(104)	(122)

Table 1
(Continued)
Data Comparing Essays With Discrepant and Comparable Scores

	LOW GROUP		HIGH GROUP		
Q5. Special Rhetorical Devices					
	Discrepant	Comparable		Discrepant	Comparable
Agree	32%	18%	Agree	34%	30%
Disagree	68%	82%	Disagree	66%	70%
	100%	100%		100%	100%
	(40)	(38)		(104)	(122)
Q6. Controversial Treatment					
	Discrepant	Comparable		Discrepant	Comparable
Agree	45%	42%	Agree	34%	30%
Disagree	55%	58%	Disagree	66%	70%
	100%	100%		100%	100%
	(40)	(38)		(104)	(122)
Q7. Stated Thesis					
	Discrepant	Comparable		Discrepant	Comparable
Agree	85%	87%	Agree	88%	90%
Disagree	15%	13%	Disagree	12%	10%
	100%	100%		100%	100%
	(40)	(38)		(104)	(122)
Q8. Adequately Developed					
	Discrepant	Comparable*		Discrepant	Comparable*
Agree	60%	32%	Agree	66%	80%
Disagree	40%	68%	Disagree	34%	20%
	100%	100%		100%	100%
	(40)	(38)		(104)	(122)

*Significant χ^2

Table 1
(Continued)
Data Comparing Essays With Discrepant and Comparable Scores

	LOW GROUP		HIGH GROUP		
Q9. Organized Paper					
	Discrepant	Comparable		Discrepant	Comparable*
Agree	72%	71%	Agree	82%	93%
Disagree	28%	29%	Disagree	18%	7%
	100%	100%		100%	100%
	(40)	(38)		(104)	(122)
Q10. Focused and Coherent Paper					
	Discrepant	Comparable		Discrepant	Comparable*
Agree	48%	50%	Agree	69%	86%
Disagree	52%	50%	Disagree	31%	14%
	100%	100%		100%	100%
	(40)	(38)		(104)	(122)
Q11. Depth of Thought					
	Discrepant	Comparable		Discrepant	Comparable
Agree	65%	55%	Agree	93%	89%
Disagree	35%	45%	Disagree	8%	11%
	100%	100%		100%	100%
	(40)	(38)		(104)	(122)
Q12. Misinformation					
	Discrepant	Comparable		Discrepant	Comparable
Agree	10%	5%	Agree	6%	2%
Disagree	90%	95%	Disagree	94%	98%
	100%	100%		100%	100%
	(40)	(38)		(104)	(122)

*Significant χ^2

Table 1
(Continued)
Data Comparing Essays With Discrepant and Comparable Scores

	LOW GROUP		HIGH GROUP		
Q13. Adequate Introduction					
	Discrepant	Comparable		Discrepant	Comparable
Agree	55%	58%	Agree	71%	73%
Disagree	45%	42%	Disagree	29%	27%
	100%	100%		100%	100%
	(40)	(38)		(104)	(122)
Q14. Adequate Conclusion					
	Discrepant	Comparable		Discrepant	Comparable
Agree	55%	45%	Agree	70%	69%
Disagree	45%	55%	Disagree	30%	31%
	100%	100%		100%	100%
	(40)	(38)		(104)	(122)
Q15. Precise Diction					
	Discrepant	Comparable		Discrepant	Comparable*
Agree	50%	53%	Agree	79%	93%
Disagree	50%	47%	Disagree	21%	7%
	100%	100%		100%	100%
	(40)	(38)		(104)	(122)
Q16. Mature Diction					
	Discrepant	Comparable		Discrepant	Comparable
Agree	65%	55%	Agree	91%	93%
Disagree	35%	45%	Disagree	9%	7%
	100%	100%		100%	100%
	(40)	(38)		(104)	(122)

*Significant x^2

Table 1
(Continued)
Data Comparing Essays With Discrepant and Comparable Scores

	LOW GROUP		HIGH GROUP	
Q17. Inflated Diction/Jargon				
	Discrepant	Comparable*	Discrepant	Comparable
Agree	12%	0%	8%	3%
Disagree	88%	100%	92%	97%
	100%	100%	100%	100%
	(40)	(38)	(104)	(122)
Q18. Varied Sentence Style				
	Discrepant	Comparable	Discrepant	Comparable
Agree	60%	42%	81%	84%
Disagree	40%	58%	19%	16%
	100%	100%	100%	100%
	(40)	(38)	(104)	(122)
Q19. Awkward Sentence Constructions				
	Discrepant	Comparable	Discrepant	Comparable*
Occas./Freq.	52%	63%	38%	25%
Never/Seldom	48%	37%	62%	75%
	100%	100%	100%	100%
	(40)	(38)	(104)	(122)
Q20. Usage Errors				
	Discrepant	Comparable	Discrepant	Comparable
Occas./Freq.	78%	74%	38%	32%
Never/Seldom	22%	26%	62%	68%
	100%	100%	100%	100%
	(40)	(38)	(104)	(122)

*Significant χ^2

Table 1
(Continued)
Data Comparing Essays With Discrepant and Comparable Scores

	LOW GROUP		HIGH GROUP		
Q21. Sentence Errors					
	Discrepant	Comparable		Discrepant	Comparable*
Occas./Freq.	52%	58%	Occas./Freq.	34%	18%
Never/Seldom	48%	42%	Never/Seldom	66%	82%
	100%	100%		100%	100%
	(40)	(38)		(104)	(122)
Q22. Dialect Errors					
	Discrepant	Comparable		Discrepant	Comparable
Occas./Freq.	35%	26%	Occas./Freq.	5%	3%
Never/Seldom	65%	74%	Never/Seldom	95%	97%
	100%	100%		100%	100%
	(40)	(38)		(104)	(122)
Q23. ESL Errors					
	Discrepant	Comparable		Discrepant	Comparable*
Occas./Freq.	22%	26%	Occas./Freq.	10%	3%
Never/Seldom	78%	74%	Never/Seldom	90%	97%
	100%	100%		100%	100%
	(40)	(38)		(104)	(122)
Q24. Spelling Errors					
	Discrepant	Comparable*		Discrepant	Comparable
Occas./Freq.	90%	71%	Occas./Freq.	57%	46%
Never/Seldom	10%	29%	Never/Seldom	43%	54%
	100%	100%		100%	100%
	(40)	(38)		(104)	(122)

*Significant χ^2

Table 1
(Continued)
Data Comparing Essays With Discrepant and Comparable Scores

	LOW GROUP		HIGH GROUP	
Q24a. Key Words Misspelled				
	Discrepant	Comparable	Discrepant	Comparable
Yes	30%	26%	8%	7%
No	70%	74%	92%	93%
	100%	100%	100%	100%
	(40)	(38)	(103)	(121)
Q24b. Simple Words Misspelled				
	Discrepant	Comparable	Discrepant	Comparable
Yes	72%	61%	43%	26%
No	28%	39%	57%	74%
	100%	100%	100%	100%
	(40)	(38)	(103)	(121)
Q24c. Troublesome Words Misspelled				
	Discrepant	Comparable	Discrepant	Comparable
Yes	40%	24%	16%	24%
No	60%	76%	84%	76%
	100%	100%	100%	100%
	(40)	(38)	(103)	(121)
Q25. Punctuation/Capitalization Errors				
	Discrepant	Comparable	Discrepant	Comparable
Occas./Freq.	98%	89%	84%	75%
Never/Seldom	2%	11%	16%	25%
	100%	100%	100%	100%
	(40)	(38)	(104)	(122)

Table 1
(Continued)
Data Comparing Essays With Discrepant and Comparable Scores

	LOW GROUP		HIGH GROUP	
Q25a. Errors in Commas	Discrepant	Comparable	Discrepant	Comparable
Yes	70%	68%	62%	60%
No	30%	32%	38%	40%
	100%	100%	100%	100%
	(40)	(38)	(103)	(121)
Q25b. Errors in Colons/Semi-Colons	Discrepant	Comparable	Discrepant	Comparable
Yes	12%	11%	18%	12%
No	88%	89%	82%	88%
	100%	100%	100%	100%
	(40)	(38)	(103)	(121)
Q25c. Errors in Quotes	Discrepant	Comparable	Discrepant	Comparable
Yes	25%	3%	16%	15%
No	75%	97%	84%	85%
	100%	100%	100%	100%
	(40)	(38)	(103)	(121)
Q25d. Errors in End Punctuation	Discrepant	Comparable	Discrepant	Comparable
Yes	25%	18%	11%	11%
No	75%	82%	89%	89%
	100%	100%	100%	100%
	(40)	(38)	(103)	(121)

Table 1
(Continued)
Data Comparing Essays With Discrepant and Comparable Scores

	LOW GROUP		HIGH GROUP	
Q25e. Errors in Apostrophes	**Discrepant**	**Comparable**	**Discrepant**	**Comparable**
Yes	20%	37%	18%	13%
No	80%	63%	82%	87%
	100%	100%	100%	100%
	(40)	(38)	(103)	(121)
Q25f. Errors in Capitalization	**Discrepant**	**Comparable**	**Discrepant**	**Comparable**
Yes	38%	37%	21%	19%
No	62%	63%	79%	81%
	100%	100%	100%	100%
	(40)	(38)	(103)	(121)
Q26. Messy Paper/Handwriting	**Discrepant**	**Comparable**	**Discrepant**	**Comparable**
Agree	15%	8%	11%	6%
Disagree	85%	92%	89%	94%
	100%	100%	100%	100%
	(40)	(38)	(104)	(122)
Q27. Adequate Length	**Discrepant**	**Comparable**	**Discrepant**	**Comparable**
Agree	80%	66%	88%	89%
Disagree	20%	34%	12%	11%
	100%	100%	100%	100%
	(40)	(38)	(104)	(122)

Table 2

2 x 2 Contingency Tables for Discrepant and Comparable Scores With Significant χ^2 for Total Group

TOTAL GROUP

Q10. Focused and Coherent Paper

	Discrepant	Comparable
Agree	91	124
Disagree	53	36

$\chi^2 = 7.49$, $p < .01$

Q15. Precise Diction

	Discrepant	Comparable
Agree	102	134
Disagree	42	26

$\chi^2 = 7.28$, $p < .01$

Q17. Inflated Diction/Jargon

	Discrepant	Comparable
Agree	13	4
Disagree	131	156

$\chi^2 = 6.12$, $p < .05$

Table 2
(Continued)
2 x 2 Contingency Tables for Discrepant and Comparable Scores
With Significant χ^2 for Total Group

TOTAL GROUP

Q21. Sentence Errors

	Discrepant	Comparable
Occas./Freq.	56	44
Never/Seldom	88	116

$\chi^2 = 4.45$, $p < .05$

Q24. Spelling Errors

	Discrepant	Comparable
Occas./Freq.	95	83
Never/Seldom	49	77

$\chi^2 = 6.21$, $p < .05$

Q25. Punctuation/Capitalization Errors

	Discrepant	Comparable
Occas./Freq.	126	125
Never/Seldom	28	35

$\chi^2 = 4.63$, $p < .05$

Table 3

2 x 2 Contingency Tables for Discrepant and Comparable Scores With Significant χ^2 for Low and High Groups

	LOW GROUP		HIGH GROUP	
Q8. Adequately Developed				
	Discrepant	Comparable	Discrepant	Comparable
Agree	24	12	69	97
Disagree	16	26	35	25
	$\chi^2 = 6.33$, $p < .05$		$\chi^2 = 4.99$, $p < .05$	
Q9. Organized Paper				
			Discrepant	Comparable
Agree			85	113
Disagree			19	9
			$\chi^2 = 6.14$, $p < .05$	
Q10. Focused and Coherent Paper				
			Discrepant	Comparable
Agree			72	105
Disagree			32	17
			$\chi^2 = 9.37$, $p < .01$	
Q15. Precise Diction				
			Discrepant	Comparable
Agree			82	114
Disagree			22	8
			$\chi^2 = 10.39$, $p < .01$	
Q17. Inflated Diction/Jargon				
	Discrepant	Comparable		
Agree	5	0		
Disagree	35	38		
	$\chi^2 = 5.08$, $p < .05$			

Table 3
(Continued)
2 x 2 Contingency Tables for Discrepant and Comparable Scores
With Significant χ^2 for Low and High Groups

LOW GROUP

HIGH GROUP

Q19. Awkward Sentence Constructions

	Discrepant	Comparable
Occas./Freq.	40	31
Never/Seldom	64	91

$\chi^2 = 4.44$, $p < .05$

Q21. Sentence Errors

	Discrepant	Comparable
Occas./Freq.	35	22
Never/Seldom	69	100

$\chi^2 = 7.26$, $p < .01$

Q23. ESL Errors

	Discrepant	Comparable
Occas./Freq.	10	4
Never/Seldom	94	118

$\chi^2 = 3.88$, $p < .05$

Q24. Spelling Errors

	Discrepant	Comparable
Occas./Freq.	36	27
Never/Seldom	4	11

$\chi^2 = 4.5$, $p < .05$